DOCUMENT RESUME

ED 467 805 TM 034 346

AUTHOR	De Champlain, Andre F.
TITLE	Assessing the Dimensionality of Simulated LSAT Item Response Matrices with Small Sample Sizes and Short Test Lengths. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-96-01
PUB DATE	1999-03-00
NOTE	20p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Admission (School); *College Entrance Examinations; *Item Response Theory; *Law Schools; Matrices; *Sample Size; Simulation; *Test Length
IDENTIFIERS	*Dimensionality (Tests); *Law School Admission Test; Type I Errors

ABSTRACT

The purpose of this study was to examine empirical Type I error rates and rejection rates for three dimensionality assessment procedures with data sets simulated to reflect short tests and small samples. The TESTFACT G superscript 2 difference test suffered from an inflated Type I error rate with unidimensional data sets, while the approximate chi squared statistic based on a NOHARM analysis did not. Rejection rates with simulated two-dimensional data sets were high for both procedures. The behavior of the G superscript 2 difference test was highly influenced by the independent variables manipulated, which was not the case for the approximate chi squared statistic. The implications of these results for small volume administrations are discussed. (Contains 5 tables and 70 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made

ED 467 805

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. VASELECK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM034346

Assessing the Dimensionality of Simulated LSAT Item Response Matrices with Small Sample Sizes and Short Test Lengths

André F. De Champlain
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-01
March 1999

LAW Services

A Publication of the Law School Admission Council

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and ***LSACD*** are trademarks of the Law School Admission Council, Inc.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Abstract 1

Introduction 1

Purpose 6

Methods 6
Unidimensional Dataset Simulations 6
Two-Dimensional Dataset Simulations 7
Analyses 9

Results 9
Unidimensional Dataset Analyses 9
Approximate χ^2 Statistic Empirical Type I Error Rates (NOHARM) 9
Approximate G^2 Difference Test Empirical Type I Error Rates (TESTFACT) 10
Multidimensional Dataset Analyses 10
Approximate χ^2 Statistic Rejection Rates (NOHARM) 11
Approximate G^2 Difference Test Rejection Rates (TESTFACT) 11

Discussion 11

References 13

BEST COPY AVAILABLE

Abstract

The assumption of unidimensionality must be met in order to legitimately use common IRT models. The validity of score-based inferences rests largely on the extent to which it can be shown that the dimensional structure underlying a test is consistent with the blueprint. Little research has been undertaken to examine the behavior of dimensionality assessment procedures in conditions similar to those encountered in small volume administrations. The purpose of this study was to examine empirical Type I error rates and rejection rates for three dimensionality assessment procedures with datasets simulated to reflect short tests and small samples. The TESTFACT G^2 difference test suffered from an inflated Type I error rate with unidimensional datasets whereas the approximate χ^2 statistic based on a NOHARM analysis did not. Rejection rates with simulated two-dimensional datasets were high for both procedures. The behavior of the G^2 difference test was highly influenced by the independent variables manipulated, which was not the case for the approximate χ^2 statistic. The implications of these results for small volume administrations are discussed.

Introduction

The assumption of unidimensionality must be met in order to legitimately use common item response theory (IRT) models. The validity of score-based inferences rests largely on the extent to which it can be shown that the dimensional structure underlying a test is consistent with the blueprint. Little research has been undertaken to examine the behavior of dimensionality assessment procedures in conditions similar to those encountered in small volume administrations. The purpose of this study was to examine empirical Type I error rates and rejections rates for three dimensionality assessment procedures with datasets simulated to reflect short tests and small samples. The TESTFACT G^2 difference test and the LISREL8 χ^2 statistic suffered from an inflated Type I error rate with unidimensional datasets, whereas the approximate χ^2 statistic based on a NOHARM analysis did not. Rejection rates with simulated two-dimensional datasets were high for all procedures. The behavior of the G^2 difference test was highly influenced by the independent variables manipulated, which was not the case for the approximate χ^2 statistic. The implications of these results for small volume administrations are discussed.

The many advantages of IRT models, namely that "sample-free" item parameter estimates and "test-free" ability estimates can be obtained, have contributed to their increased use in education and psychology to address a multitude of measurement-related issues. Recently, IRT models have also been popular and quite useful with respect to the development of computerized adaptive tests (CAT; Hambleton, Zaal, & Pieters, 1993; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990). Law School Admission Council (LSAC) staff currently employ an IRT model to estimate the statistical characteristics of test items and equate scores obtained on alternate forms of a test as well as to assemble new forms. However, in order to legitimately use common IRT models, several strict assumptions must be met, one of which is unidimensionality of the latent ability space. It is assumed, when using most IRT models, that the probability of a correct response on a given item requires a single underlying latent trait, often interpreted as a proficiency or ability being measured by the test. For example, the probability of a correct response on a given item using the three-parameter logistic IRT function (Lord & Novick, 1968) is given by

$$P(x=1 \mid a,b,c,\theta) = c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}; \quad (1)$$

that is, the probability of correctly answering the item (denoted by $x = 1$) is assumed to be dependent upon an item discrimination (a), difficulty (b), and lower asymptote (c) parameter as well as the latent trait or proficiency (θ) postulated to underlie the item responses. It is clear that the assumption of unidimensionality is often violated with actual achievement datasets where the response to an item is dependent upon not only the hypothesized proficiency but also several other secondary abilities. For example, the dependencies that exist between item sets in the Analytical Reasoning (AR) and Reasoning Comprehension (RC) sections of the Law School Admission Test (LSAT), due to the presence of passages, contribute in increasing their dimensional complexity to include factors other than the proficiency hypothesized to underlie the item responses (i.e., AR and RC abilities).

This led researchers to propose a multitude of descriptive statistics to assess dimensionality, or more commonly, departure from the assumption of unidimensionality. Table 1 presents some of the procedures proposed thus far in the literature along with their respective contributors.

TABLE 1

Procedures proposed for assessing the dimensionality of a set of item responses

Procedures	References
Indices based on linear factor analysis/principal component analysis	Berger & Knol (1990) Collins, Cliff, McCormick, & Zatkin (1986) De Ayala & Hertzog (1989) Hambleton & Rovinelli (1986) Hattie (1984, 1985) Nandakumar (1994) Reckase (1979) Zwick & Velicer (1986)
Nonmetric multidimensional scaling	De Ayala & Hertzog (1989) Jones (1988) Jones, Sabers, & Trosset (1987) Koch (1983) Reckase (1981)
Tucker's procedure for assessing dimensionality	Roznowski, Tucker, & Humphreys (1991)
Humphrey's procedure for assessing dimensionality	Roznowski, Tucker, & Humphreys (1991)
Modified parallel analysis	Ben-Simon & Cohen (1990) Budescu, Cohen, & Ben-Simon (1994) Drasgow & Lissak (1983) Hulin, Drasgow, & Parsons (1983)
Bejar's dimensionality assessment procedure	Bejar (1980, 1988) Hambleton & Rovinelli (1986) Kingsbury (1985) Liou (1988)
The Holland-Rosenbaum procedure	Ben-Simon & Cohen (1990) Holland (1981) Holland & Rosenbaum (1986) Nandakumar (1994) Rosenbaum (1984) Zwick (1987)
Stout's essential dimensionality procedure	De Champlain (1992) De Champlain & Tang (1993) Gessaroli & De Champlain (1996) Junker & Stout (1994) Nandakumar (1987, 1991, 1994) Nandakumar & Stout (1993) Stout (1987, 1990)
Indices and statistics based on full-information nonlinear factor analysis	Berger & Knol (1990) Bock, Gibbons, & Muraki (1988) Dorans & Lawrence (1988) Kingston (1986) Kingston & McKinley (1988) Morgan (1989) Muraki & Engelhard (1985)
Indices and statistics based on limited-information nonlinear factor analysis	Berger & Knol (1990) De Champlain (1992) De Champlain & Gessaroli (1991) De Champlain & Tang (1993) Gessaroli & De Champlain (1996) Hambleton & Rovinelli (1986) Hattie (1984, 1985) Knol & Berger (1991) Nandakumar (1994)

At the present time, Stout's DIMTEST procedure and indices, as well as statistics based on nonlinear factor analysis (NLFA) appear to be the two most popular and promising procedures for assessing the dimensionality of a set of item responses.

Stout proposed a nonparametric procedure (the *T* statistic) that is based on his concepts of *essential independence* and *essential dimensionality* (Nandakumar, 1991; Nandakumar & Stout, 1993; Stout, 1987, 1990). Stout, Junker, Nandakumar, Chang, and Steidinger (1991) developed the computer program DIMTEST to estimate the value of the *T* statistic for any given dataset. Essential dimensionality can be defined as the number of latent traits that is needed to satisfy the assumption of essential independence given by

$$\frac{1}{n(n-1)} \sum_{1 \le i \le j \le n} | Cov(U_i, U_j | \theta) | \simeq 0 \quad n \to \infty ; \tag{2}$$

that is, a mean absolute residual covariance value that tends towards zero at fixed latent trait levels as the number of items increases towards infinity. The terms shown in Equation 2 can be defined as follows:

n = the number of items;
U_i = the response to item i for a randomly selected test taker; and
U_j = the response to item j for a randomly selected test taker.

Several versions of the *T* statistic have been proposed by Stout (1987, 1990) and Nandakumar and Stout (1993) to test the assumption of essential unidimensionality (d_e) given by

$$H_0 : d_e = 1$$

$$H_a : d_e > 1$$

where d_e corresponds to the number of dimensions required to satisfy the assumption of essential independence. The first step involved in computing the *T* statistic entails dividing a set of items into two distinct subsets, labelled *AT1* and *AT2*, and a partitioning test or *PT*. The *AT1* items are selected as the unidimensional subset, generally based on the factor loadings estimated after fitting a linear factor analytic model to the tetrachoric item correlation matrix. The *AT2* items are chosen to correct for bias, which results from matching test takers based on their number-right score on the remaining items; that is, the *PT* test. Nandakumar and Stout (1993) recommend using the T_2 version of the statistic in most instances given its demonstrated low Type I error and high power. The T_2 statistic can be defined as follows:

$$T_2 = \frac{T_{L,2} - T_b}{\sqrt{2}} \tag{3}$$

where

$$T_{L,2} = \frac{1}{K^{1/2}} \left(\sum_{k=1}^{K} \frac{X_k}{S_k^2} \right) \tag{4}$$

and

K = the number of subgroups based on the *PT* item subscore;
k_i = the i_{th} subgroup of test takers based on the *PT* item subscore; and
S_k = the standard error of the T_2 statistic.

Note that the T_b statistic is identical to the $T_{L,2}$ with the exception that it is computed for *AT2* items. Readers interested in obtaining more information regarding the computation of the T_2 statistic should consult Nandakumar and Stout (1993). The T_2 statistic is asymptotically normally distributed with a mean and

BEST COPY AVAILABLE

standard deviation equal to zero and one respectively, under the null hypothesis of unidimensionality. Nandakumar and Stout (1993) showed, in a series of Monte Carlo studies, that the T_2 statistic was generally accurate in correctly determining essential unidimensionality or violation of the assumption with multidimensional datasets, except when the test contained few items (less than 25) and the sample sizes were small (less than 750 test takers). Consequently, the procedure cannot be used in many instances, for example, with CAT forms, where short test lengths and small sample sizes are a common occurrence due in part to the assembly algorithms used and the "on-demand" nature of the scheduling.

Another promising approach, with respect to assessing the dimensionality of an item response matrix, is the one that treats common IRT models as a special case of a more general NLFA model. Bartholomew (1983), Goldstein and Wood (1989), McDonald (1967), and Takane and De Leeuw (1987), to name a few, have shown that common IRT models and NLFA models are mathematically equivalent. This led other researchers to suggest that a useful way of assessing the dimensionality of a set of item responses might entail analyzing the residual correlation or covariance matrix obtained after fitting an m-factor model to an item response matrix, where m corresponds to the number of factors or dimensions. The rationale underlying this approach is as follows: zero residual correlations obtained after fitting a unidimensional (i.e., one-factor) model to an item response matrix would be indicative of unidimensionality. A host of descriptive indices and hypothesis tests have been proposed to assess dimensionality based on both *limited-information* and *full-information* NLFA models (see Hattie, 1984, 1985 for a review of earlier indices). The estimation of parameters in limited-information NLFA models is restricted to the information contained in the lower-order marginals (e.g., the pairwise relationships between items) whereas the information included in all higher-order relationships (i.e., in the item response vectors) is utilized to estimate the parameters of full-information NLFA models.

Gessaroli and De Champlain (1996) investigated the usefulness of an approximate chi-square statistic for the assessment of dimensionality that is based on the estimation of parameters for a limited-information m-factor model using the *polynomial approximation to a normal ogive model* (McDonald, 1967), as implemented in the computer program NOHARM (Fraser & McDonald, 1988). This approximate chi-square statistic, originally proposed by Bartlett (1950) and outlined in Steiger (1980a, 1980b), tests the null hypothesis that the off-diagonal elements of a residual correlation matrix are equal to zero after fitting an m-factor NLFA model and can be defined as

$$\chi^2 = (N-3)\sum_{i=1}^{k}\sum_{j=1}^{i-1} Z_{ij}^{2(r)}, \qquad (5)$$

where $Z^2_{ij}(r)$ is the square of the Fisher Z corresponding to the residual correlation between items i and j ($i, j = 1, ..., k$) and N is the number of test takers. Under the null hypothesis of unidimensionality, this statistic is distributed approximately as a central chi-square with $df = .5k\,(k - 1) - t$, where k is the number of items and t is the total number of parameters estimated in the NLFA model. Although the approximate χ^2 statistic is based on unweighted least-squares estimation (ULS), and hence is weak in its theoretical foundation, Browne (1977, 1986) has indicated that the latter statistic is often equivalent to a χ^2 obtained from generalized least-squares estimation (GLS). Browne states that, in most instances, χ^2 statistics based on ULS and GLS tend to differ only slightly. Therefore, the approximate χ^2 statistic outlined in Equation 5 has the potential of being a useful practical tool for the assessment of dimensionality. Simulation and real data studies (De Champlain, 1996; Gessaroli & De Champlain, 1996) have shown that the Type I error rate for the approximate chi-square statistic tends to be at or below nominal alpha levels. With multidimensional datasets, rejection rates were generally high, even in some instances with datasets containing as few as 15 items and 500 test takers, which was not the case for the T statistic (De Champlain, 1992; Gessaroli & De Champlain, 1996). The computer program TESTFACT (Wilson, Wood, & Gibbons, 1991) allows the user, among other things, to estimate the parameters and the fit of various full-information factor analytic models using the marginal maximum likelihood (MML) procedure outlined by Bock and Aitkin (1981) via the EM algorithm of

Dempster, Laird, and Rubin (1977). The thresholds and factor loadings included in the model are estimated so as to maximize the following multinomial probability function:

$$L_m = P(X) = \frac{N!}{r_1! r_2! \ldots r_s!} \tilde{P}_1^{r_1} \tilde{P}_2^{r_2} \ldots \tilde{P}_s^{r_s}, \quad (6)$$

where r_s is the frequency of response pattern s and $\tilde{P}_s$ is the marginal probability of the response pattern based on the item parameter estimates. The function given in Equation 6 is customarily referred to as *full-information item factor analysis* (Bock, Gibbons, & Muraki, 1988). The user can also assess the fit of a given full-information factor analytic model using a likelihood-ratio chi-square statistic that is provided in TESTFACT. This statistic can be defined as

$$G^2 = 2 \sum_{1}^{2^n} r_l \, ln \frac{r_l}{N \tilde{P}_l}, \quad (7)$$

where r_l is the frequency of response vector l and $\tilde{P}_l$ is the probability of response vector l. The degrees of freedom for this statistic are equal to

$$2^n(m + 1) + m(m - 1)/2,$$

where n is the number of items and m is the number of factors. However, Mislevy (1986) has indicated that this G^2 statistic often poorly approximates the chi-square distribution given the large number of empty cells typically encountered with actual datasets (the number of unique response vectors is equal to 2^n). Hence, Haberman (1977) recommends using a likelihood-ratio chi-square difference test to assess the fit of alternative models. The G^2 difference test is computed in the following fashion:

$$G^2_{diff} = G^2_{1-F} - G^2_{2-F}, \quad (8)$$

where $G^2_{1\text{-}F}$ is the value of the likelihood-ratio chi-square statistic obtained after fitting a one-factor model (c.f. Equation 7) and $G^2_{2\text{-}F}$ is the value of the likelihood-ratio chi-square statistic obtained after fitting a two-factor model. The degrees of freedom for the difference test are also computed by subtracting those associated with the one- and two-factor model fit statistics.

However, preliminary research has shown that the likelihood-ratio chi-square difference test is generally unable to correctly identify the number of dimensions underlying an item response matrix (Berger & Knol, 1990). However, the small number of replications (10) performed in the latter study limits the extent to which these results can be generalized to other conditions.

Although these fit statistics have been shown to be useful and informative for the assessment of dimensionality, few studies have examined their behavior with small sample sizes and short test lengths. This type of study seems imperative given the current interest in CAT by the LSAC as evidenced by the current five-year research project on this subject. Dimensionality assessment is especially critical within a CAT environment where several test forms are "tailored" to different test takers according to their ability level. These CAT forms should be comparable with respect to their dimensional structure in order to ensure valid score-based inferences for all test takers, irrespective of the set of items administered. The assessment of dimensionality is also critical within a computerized mastery testing (CMT) setting where a small set of items is typically administered to all test takers in the first stage of testing in order to determine whether test takers can be clearly categorized as masters/nonmasters or whether further sets of items need to be given before making any final decision as to their status. The first subset of items administered within this multistage or sequential design often contains very few items. Hence, it is critical to ascertain whether the dimensional structure of this initial test is consistent with that of subsequent subtests in order to ensure that the design is fair for all test takers, regardless of their ability level.

BEST COPY AVAILABLE

Purpose

The purpose of this study is two-fold: (1) examine the empirical Type I error rates calculated for the approximate chi-square statistic and the likelihood-ratio chi-square difference test with unidimensional datasets simulated to vary according to test length and sample size; and (2) examine the rejection rates obtained for the approximate chi-square statistic and the likelihood-ratio chi-square difference test with two-dimensional item response matrices generated to vary as a function of sample size, test length, and degree of correlation between the latent traits.

Methods

Unidimensional Dataset Simulations

Dichotomous unidimensional item response vectors were simulated according to the three-parameter logistic IRT function outlined in Equation 1 in the first part of this study. Data sets were generated to vary according to two different test lengths (20 and 40 items) as well as three sample sizes (250, 500, and 1,000 test takers). Note that the simulated 40-item datasets were composed of two 20-item tests; that is, the item parameters utilized to simulate responses to items 21-40 were identical to those selected to generate responses to items 1-20. In order to simulate item responses that are typical of those encountered at the LSAC, 20 IRT-item parameters were randomly selected from one form of the LSAT and used in the item response generation process. The item parameters that were chosen to simulate unidimensional item response vectors are shown in Table 2.

TABLE 2

True unidimensional item parameters

Item	*a*	*b*	*c*
1	0.622132	-1.710310	0.119606
2	0.779642	0.470174	0.079124
3	0.806952	0.161454	0.162809
4	0.842712	0.081694	0.140943
5	1.152409	1.679257	0.153869
6	0.558630	-1.387155	0.119606
7	0.341596	-0.599501	0.119606
8	0.878353	1.081976	0.058036
9	0.957605	0.916684	0.196364
10	1.086517	0.693614	0.042316
11	0.751002	-0.696663	0.119606
12	0.551905	-0.315874	0.119606
13	0.630988	1.696784	0.223633
14	0.552291	-1.294931	0.119606
15	0.785618	-0.285280	0.095973
16	0.730466	-0.402966	0.119606
17	0.845300	0.004327	0.188632
18	0.792140	1.138772	0.155819
19	0.822973	1.540107	0.073885
20	0.601753	1.358651	0.111348

Latent trait values were also simulated according to an $N(0,1)$ distribution. Each cell of this 2 (test length) x 3 (sample size) design was replicated 100 times for a total of 600 unidimensional datasets.

The fit of a unidimensional model was then ascertained for each of the 600 unidimensional datasets using both TESTFACT (Wilson, Wood, & Gibbons, 1991) as well as NOHARM (Fraser & McDonald, 1988).

More precisely, one- and two-factor models were fit to each simulated unidimensional dataset with TESTFACT using all default values. As mentioned previously, the likelihood-ratio chi-square difference test was selected as the fit statistic for all unidimensional dataset analyses given that it follows a chi-square distribution even in the presence of sparse frequency tables (Haberman, 1977). Again, the G^2 difference test is obtained by simply subtracting the G^2 value obtained after fitting a two-factor model from that computed after fitting a unidimensional model.

The fit of a unidimensional model (i.e., one-factor model) was also ascertained using NOHARM (Fraser & McDonald, 1988). The approximate χ^2 statistic was then computed for each dataset using the computer program CHIDIM (De Champlain & Tang, 1997).

Two-Dimensional Dataset Simulations

In the second part of the study, dichotomous two-dimensional item response vectors were simulated based on a multidimensional extension of the three-parameter logistic IRT model (M3PL; Reckase, 1985) outlined in Equation 1. The probability of a correct response on item i (denoted by $x = 1$), based on this compensatory M3PL model, is given by

$$P_i(x_i = 1 \mid a_i, d_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j + d_i)}}{1 + e^{a_i(\theta_j + d_i)}}, \tag{9}$$

where

a_i = a vector of discrimination parameters for item i;
d_i = a scalar parameter related to the difficulty of item i; and
θ_j = a latent trait vector.

Reckase (1985) states that a multidimensional item discrimination parameter (MDISC) can be estimated using the following equation:

$$MDISC_i = \sqrt{\sum_{k=1}^{n} a_{ik}^2}, \tag{10}$$

where a_{ik} is the discrimination parameter of item i on dimension k (k = 1,2, ..., l). In a similar fashion, the multidimensional item difficulty (MDIF) for item i can also be computed using the following formula:

$$MDIF_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{n} a_{ik}^2}}. \tag{11}$$

It should be noted that Reckase (1985) recommends providing direction cosines in addition to the distance outlined in Equation 11 when describing the MDIF value of an item. However, he does suggest that the distance parameter can be interpreted much like a b parameter would be for a unidimensional logistic IRT model.Past research undertaken to assess the dimensionality of the LSAT has convincingly shown that a two-factor model appears to adequately account for the item response probabilities estimated on several forms of the test (Ackerman, 1994; Camilli, Wang, & Fesq, 1995; De Champlain, 1996; Roussos & Stout, 1994).

BEST COPY AVAILABLE

The first dimension, categorized as *deductive reasoning*, loads on AR items while the second factor, which loads on Logical Reasoning (LR) and RC items, has been labelled as *reading/informal reasoning*. Approximately 25% of the items on any given form of the LSAT measure this deductive reasoning skill whereas the remaining 75% of the items require reading/informal reasoning. As was the case with unidimensional datasets, an item parameter structure that resembles that found on a typical form of the LSAT was selected in order to generate more "authentic" item responses. More precisely, the first dimension (factor) was constrained to load on 25% of the items while the probability of a correct response on the remaining 75% of the items was solely a function of the second latent trait. As well, (unidimensional) item discrimination parameters were randomly selected from actually administered LSAT AR + LR/RC items and used to simulate the first and second dimensions in this study. The unidimensional item difficulty parameter estimates for these items were treated as MDIF values in the simulations. The item parameters utilized in the two-dimensional simulations are shown in Table 3.

TABLE 3

True two-dimensional item parameters

Item	a_1	a_2	*MDIF*	*c*
1	0.622132	0.000000	-1.710310	0.119606
2	0.806592	0.000000	0.161454	0.162809
3	0.842712	0.000000	0.081694	0.140943
4	0.882054	0.000000	0.854201	0.184434
5	0.904691	0.000000	1.371124	0.242642
6	0.000000	0.644494	-0.892373	0.119606
7	0.000000	0.878353	1.081976	0.058036
8	0.000000	0.957605	0.916684	0.196364
9	0.000000	0.946642	1.520134	0.224578
10	0.000000	0.803943	-1.139963	0.119606
11	0.000000	0.751002	-0.696663	0.119606
12	0.000000	0.551905	-0.315874	0.119606
13	0.000000	0.688839	0.632910	0.145847
14	0.000000	0.808383	0.554415	0.208314
15	0.000000	0.567085	-0.087459	0.119606
16	0.000000	0.783265	0.256477	0.206116
17	0.000000	0.694929	-1.357711	0.119606
18	0.000000	0.543069	-0.608002	0.119606
19	0.000000	0.792140	1.138772	0.155819
20	0.000000	0.773915	0.280484	0.246003

In addition, the two-dimensional datasets were generated to vary as a function of the same two test lengths (20 and 40 items) and three sample sizes (250, 500, and 1,000 test takers) previously outlined in the unidimensional conditions. The 40-item datasets were also composed of two 20-item tests. Also, past research has shown that the correlation between reading/informal reasoning and deductive reasoning proficiencies on a large number of LSAT forms is at or near 0.70 (Camilli, Wang, & Fesq, 1995; De Champlain, 1996). Hence, the correlation between both latent traits was set at either 0.00 or 0.70 in the two-dimensional simulations. Finally, each cell of this 2 (test length) x 3 (sample size) x 2 (latent trait correlation) design was replicated 100 times for a total of 1,200 two-dimensional datasets.

Also, the fit of a one- versus a two-factor full-information factor analytic model was assessed using the likelihood-ratio chi-square difference test provided in TESTFACT. In addition, the fit of a unidimensional model was ascertained with the approximate χ^2 statistic, computed after fitting a one-factor model to each two-dimensional item response matrix with the computer program NOHARM (Fraser & McDonald, 1988).

Analyses

In order to investigate the effects of the independent variables on the empirical Type I error rates and rejection rates, separate logit-linear analyses were performed for the approximate χ^2 and the likelihood-ratio chi-square difference test for each of the unidimensional and multidimensional conditions. Specifically, logit-linear analyses were performed with the objective of fitting the most parsimonious model to the response frequencies. With respect to unidimensional datasets, the independent variables were test length and sample size while the dependent variable was the number of acceptances and rejections of the null hypothesis. This variable was labelled "rejection decision." The logit-linear analyses were done in a forward hierarchical manner; that is, starting with the simplest main effect and then fitting incrementally more complex models while adhering to the principle that lower-order effects are also included in the model. The likelihood-ratio χ^2 was employed as the fit statistic. A model was deemed to be acceptable if the corresponding *p*-value was equal to or greater than 0.15. Any individual effect was considered to be significant if the size of the absolute z-value was greater than 2.0. With regards to simulated two-dimensional datasets, the independent variables were test length, sample size, and latent trait correlation whereas the dependent variable was rejection decision. Results are presented for the simulated unidimensional and multidimensional datasets separately. It should be noted that, for the sake of simplicity, associations will be presented with respect to the impact of the independent variable(s) only. For example, if the test length by rejection decision association was significant, it would be referred to as the effect of test length.

Results

Unidimensional Dataset Analyses

The number of false rejections of the assumption of unidimensionality based on the 100 datasets for each of the simulated conditions are shown in Table 4.

TABLE 4

Rejections of unidimensionality per 100 trials for unidimensional datasets (nominal ∞ = .05)

	Fit Statistic			
	Approximate χ^2 (NOHARM)		G^2 Difference Test (TESTFACT)	
Test Length	20 items	40 items	20 items	40 items
Sample size				
250	0	1	58	79
500	0	0	41	77
1,000	5	7	17	77

Approximate χ^2 Statistic Empirical Type I Error Rates (NOHARM)

The empirical Type I error rates tended to be below or near the nominal ∞ level (.05). In fact, the maximum number of rejections of the assumption of unidimensionality in any given condition was 7/100 for datasets simulated to contain 40 items and 1,000 test takers. Logit-linear analyses results show that a model including sample size as the sole independent variable was sufficient in adequately accounting for the frequency of rejections (and acceptances) of the assumption of unidimensionality,

$$L^2 (4) = 1.75, p = .782.$$

BEST COPY AVAILABLE

The effect of sample size was quite clear. There was only one false rejection (.005) of the assumption of unidimensionality for datasets simulated to include 250 test takers and none for item response matrices generated to contain 500 test takers. However, the assumption of unidimensionality was incorrectly rejected for 12 (.06) datasets simulated to contain 1,000 test takers.

Approximate G^2 Difference Test Empirical Type I Error Rates (TESTFACT)

The number of incorrect rejections of the assumption of unidimensionality was quite large in all simulated conditions when based on the likelihood-ratio chi-square difference test provided in TESTFACT. Empirical Type I error rates ranged from 0.17 (for datasets that included 20 items and 1,000 test takers) to .79 (for datasets that contained 40 items and 250 test takers). These results are clearly indicative of a severe inflated Type I error rate problem when using the G^2 difference test to determine whether an item response matrix is unidimensional or not, at least with datasets similar to those simulated in the present study. The results obtained from the logit-linear analyses indicate that a fully saturated model, including the main effects of test length and sample size as well as the interaction of both variables, is required to adequately explain the frequencies of rejection and acceptance rates, L^2 (0) = 0.00, p = 1.00. All of these effects had absolute z-values greater than or equal to 2.0.

As is traditionally the case, the effects of the independent variables found in the higher-order interaction will first be explained. Results show that the number of false acceptances of the assumption of unidimensionality decreased sharply for 20-item datasets from 58/100 rejections for item response matrices simulated to contain 250 test takers to 41/100 rejections for 500 test taker datasets and finally, 17/100 rejections for datasets simulated to include 1,000 test takers. However, this drop in empirical Type I error rates was absent for the 40-item datasets. For the latter datasets, empirical Type I error rates remained quite constant across the three sample sizes. The empirical Type I error rates were equal to .79, .77, and .77 for 40-item datasets simulated to contain 250, 500, and 1,000 test takers respectively.

Multidimensional Dataset Analyses

The number of rejections of the assumption of unidimensionality based on the 100 datasets for each of the simulated two-dimensional conditions are shown in Table 5.

TABLE 5

Rejections of unidimensionality per 100 trials for two-dimensional datasets (nominal ∞ = .05)

		Fit Statistic			
		Approximate χ^2 (NOHARM)		G^2 Difference Test (TESTFACT)	
Latent Trait Correlation	Test Length	20 items	40 items	20 items	40 items
	Sample Size				
$r\,\theta_1\,\theta_2$ = 0.00	250	100	100	100	100
	500	100	100	100	100
	1,000	100	100	100	100
$r\,\theta_1\,\theta_2$ = 0.70	250	99	100	77	96
	500	100	100	79	97
	1,000	100	100	94	99

Approximate χ^2 Statistic Rejection Rates (NOHARM)

Results clearly show that the approximate χ^2 statistic was able to consistently identify the (true) multidimensional nature of the simulated datasets. The assumption of unidimensionality was rejected for 1,199/1,200 (99.9%) simulated datasets. Not surprisingly, the logit-linear analyses results indicate that a model including only the dependent variable rejection decision was sufficient to explain the observed frequencies, L^2 (11) = 4.97, p = 0.932. Neither test length, sample size, nor latent trait correlation had a significant effect on the probability of rejecting the assumption of unidimensionality when based upon the approximate χ^2 statistic.

Approximate G^2 Difference Test Rejection Rates (TESTFACT)

There was a considerably greater degree of variability in rejection rates based on the full-information factor analyses. Rejection rates ranged from 77/100 (20-item datasets simulated to contain 250 test takers and to reflect zero correlation between latent traits) to 100/100 (all conditions that specified zero correlation between the two latent traits). Logit-linear analyses results yielded a model that included the main effects of test length and sample size as well as the latent trait correlation, L^2 (11) = 0.087, p = 1.00.

With respect to the main effect of test length, results indicate that the number of failures to reject unidimensionality decreased significantly from the 20-item datasets (50/600 or 0.083 false acceptances of unidimensionality) to the 40-item datasets (8/600 or .013 false acceptances of unidimensionality). Regarding the main effect of sample size, results show that the number of false acceptances of the assumption of unidimensionality remained fairly stable for the 250 and 500 test taker datasets (respectively, 27/400 or 0.067 false acceptances and 24/400 or 0.06 false acceptances of unidimensionality) but dropped noticeably for datasets that contained 1,000 test takers (7/400 or 0.017 false acceptances of unidimensionality). Finally, with respect to the latent trait correlation main effect, findings indicate that the number of false acceptances of the assumption of unidimensionality increased drastically from 0/600 for datasets simulated to have zero correlation between both proficiencies to 58/600 (0.097) for item response matrices generated to reflect a correlation of 0.7 between both latent traits.

Discussion

The use of indices and statistics based on NLFA has become increasingly popular as a means of assessing the dimensionality of an item response matrix. Indices and statistics based on both limited- and full-information factor analytic models are currently available to the practitioner interested in determining the number of dimensions underlying a set of item responses. Although these indices have been shown to be useful and accurate in many testing conditions, few studies have investigated the behavior of these procedures with small sample sizes and short tests; that is, conditions that are typically encountered within CAT and CMT frameworks. Therefore, the purpose of this investigation was to compare the empirical Type I error rates and rejection rates obtained using two NLFA fit statistics with conditions simulated to contain short tests and small sample sizes. More precisely, the behavior of an approximate χ^2 statistic (Gessaroli & De Champlain, 1996) based on McDonald's (1967) limited-information NLFA model as well as a likelihood-ratio G^2 difference test based on Bock, Gibbons, and Muraki's (1988) full-information item factor analytic model, were examined.

With respect to empirical Type I error rates, results show that the G^2 difference test suffers from a severe inflated Type I error rate problem, irrespective of the condition simulated. In addition, the interaction of both independent variables manipulated (i.e., sample size and test length) appears to be related to the probability of correctly accepting or incorrectly rejecting the assumption of unidimensionality. The approximate χ^2 statistic, on the other hand, had empirical Type I error rates that were below or near the nominal α level (.05) in all conditions. However, it is important to point out that the probability of accepting or rejecting the assumption of unidimensionality, when based upon the latter statistic, was dependent upon sample size. This result is not surprising given that the probability of rejecting a model of restricted dimensionality is often dependent upon sample size with chi-square distributed statistics (Marsh, Balla, & McDonald, 1988).

Regarding rejection rates with (true) two-dimensional datasets, findings again show that all independent variables manipulated; that is, test length, sample size, and latent trait correlation, had a significant effect on the probability of rejecting the assumption of unidimensionality based on the G^2 difference test. Although rejection rates were generally acceptable (varying from 77/100 to 100/100 datasets), it is important to point out that this high level of power is more than likely attributable to the inflated Type I error rates previously reported with the simulated unidimensional datasets. On the other hand, the approximate χ^2 statistic, based on a NOHARM analysis, was able to correctly reject the assumption of unidimensionality for all but one of the two-dimensional simulated datasets. In addition, none of the independent variables had an effect on the probability of correctly rejecting (or incorrectly accepting) the assumption of unidimensionality.

Mood, Graybill, and Boes (1974) state that a statistical test which displays a small Type I error rate (ideally 0) as well as a high probability of rejecting a false null hypothesis (ideally unity) is worthy of merit. The results obtained in this study would seem to suggest that the approximate χ^2 statistic possesses these desirable qualities, at least for the conditions simulated. Also, Roznowski, Tucker, and Humphreys (1991) suggest that practitioners should strive to select dimensionality assessment indices that are "robust to changes in levels of parameters and lack substantial interaction among parameters" (p.124). Although the empirical Type I error rates obtained with the approximate χ^2 statistic were affected by sample size, none of the manipulated variables significantly impacted upon its rejection rates with two-dimensional datasets. On the other hand, both empirical Type I error rates and rejection rates based on the G^2 difference test were highly dependent upon test length, sample size, and latent trait correlation (with two-dimensional datasets).

In summary, the preliminary findings reported with respect to the approximate χ^2 statistic were encouraging for the following reasons:

- the procedure appears to have low Type I error rates (below or near the nominal level);
- rejection rates were very high with two-dimensional datasets; and
- the statistic was relatively unaffected by the sample sizes, test lengths, and latent trait correlation levels simulated.

However, it is important to emphasize that these findings are preliminary and that caution should be exercised when interpreting, and especially, generalizing results to other conditions. Therefore, it is important to underscore the limitations associated with this investigation as well as offer suggestions for future research in this area.

First and foremost, the conditions that were simulated in the present study reflect *some* of the dataset features that might be encountered within a CAT and CMT framework. Obviously, there are a multitude of factors, in addition to small item sets and small samples, that contribute to making CAT and CMT forms so uniquely distinct from their paper-and-pencil counterparts. For example, context effects, attributable to the large number of "tailored" forms administered at any given time, are prevalent in CAT and CMT forms. The inclusion of this factor in future studies examining the behavior of dimensionality assessment procedures should be of the utmost importance.

Second, it is important to point out that NOHARM does not estimate latent trait values but rather assumes that they are distributed ~ $N(0,1)$. TESTFACT, on the other hand, does estimate proficiency scores for all test takers. Given that the latent trait values in this study were simulated according to a standard normal distribution (i.e., that conform exactly to the NOHARM assumption), this could have advantaged the approximate χ^2 and partially account for its superior performance over the G^2 difference test provided in TESTFACT. Nonetheless, preliminary findings showed that the empirical Type I error rates computed for the approximate χ^2 were not severely affected with certain nonnormal latent trait distributions (De Champlain & Tang, 1993). However, more research needs to be undertaken to assess the performance of the approximate χ^2 statistic in a larger number of conditions, including under various proficiency distributions, before making any definite conclusions as to its usefulness in assessing dimensionality with datasets containing few items and small samples.

Third, it is important to point out that the fit of a simple (unidimensional) model was examined for all simulated datasets. The fit of more complex models (e.g., two-, three-dimensional models) should also be part of any future investigations so as to determine whether the approximate χ^2 statistic and the G^2 difference test are able to identify the (true) number of dimensions underlying item response matrices.

Finally, it is important to mention that only two procedures were examined in this study. Given the large number of indices and statistics proposed for the assessment of dimensionality (c.f. Table 1), it would seem imperative to undertake a comparative study that would allow the respective strengths and weaknesses of each approach to be highlighted.

Hopefully, the results presented in this study will offer some information to practitioners interested in using either the approximate χ^2 statistic or the G^2 difference test for assessing the dimensionality of datasets that contain few items and small samples. Also, it is hoped that these findings will foster future research in this area and eventually lead to helpful guidelines with respect to the assessment of dimensionality of LSAT forms administered within a CAT framework.

References

Ackerman, T. (1994, April). *Graphical representation of multidimensional IRT analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics, 22*, 229-243.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77-85.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bejar, I. I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12*, 377-379.

Ben-Simon, A. & Cohen, Y. (1990, April). *Rosenbaum's test of unidimensionality: Sensitivity analysis*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Berger, M. P. F., & Knol, D. L. (1990, April). *On the assessment of dimensionality in multidimensional item response theory models*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Bock, D. R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 4*, 443-459.

Bock, D. R., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Browne, M. W. (1977). The analysis of patterned correlation matrices by generalized least-squares. *British Journal of Mathematical and Statistical Psychology, 30*, 113-124.

Browne, M. W. (1986). *Robustness of statistical inference in factor analysis and related models* (Research Report 86-1). Pretoria, South Africa: University of South Africa, Department of Statistics.

Budescu, D. V., Cohen, Y., & Ben-Simon, A. (1994, April). *A revised modified parallel analysis (RMPA) for the construction of unidimensional item pools*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79-96.

BEST COPY AVAILABLE

Collins, L. M., Cliff, N., McCormick, D. J., & Zatkin, J. L. (1986). Factor recovery in binary datasets: A simulation. *Multivariate Behavioral Research, 21,* 377-391.

De Ayala, R. J., & Hertzog, M. A. (1989, March). *A comparison of methods for assessing dimensionality for use in Item Response Theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

De Champlain, A. (1992). *Assessing test dimensionality using two approximate chi-square statistics.* Unpublished doctoral dissertation, University of Ottawa, Ottawa, Ontario, Canada.

De Champlain, A. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33,* 181-201.

De Champlain, A., & Gessaroli, M. E. (1991, April). *Assessing test dimensionality using an index based on nonlinear factor analysis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

De Champlain, A., & Tang, K.L. (1993, April). *The effect of nonnormal ability distributions on the assessment of dimensionality.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

De Champlain, A., & Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57,* 174-178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Dorans, N. J., & Lawrence, I. M. (1988, April). *An item parcel approach to assessing the dimensionality of test data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68,* 363-373.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Gessaroli, M. E. & De Champlain, Al (1996). Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33,* 157-179.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139-167.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics, 5,* 1148-1169.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1993). Computerized adaptive testing: Theory, applications and standards. In R. K. Hambleton and J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 341-366). Boston: Kluwner Academic Publishers.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523-1543.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow-Jones Irwin Publishing Company.

Jones, P. B. (1988, April). *Assessment of dimensionality in dichotomously-scored data using multidimensional scaling: Analysis of HSMB data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Jones, P. B., Sabers, D. L., & Trosset, M. (1987). *Dimensionality assessment for dichotomously scored items using multidimensional scaling* (Report No. TM 870 416). Tucson, AZ: University of Arizona. (ERIC Document Reproduction Service No. ED 283 877).

Junker, B. W., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B .D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp 31-36). Ottawa, ON: University of Ottawa, Edumetrics Research Group.

Kingsbury, G. G. (1985). *A comparison of item response theory procedures for assessing response dimensionality* (Report No. TM 850 477). Portland, OR: Portland Public Schools. (ERIC Document Reproduction Service No. ED 261 075).

Kingston, N. (1986). *Assessing the dimensionality of the GMAT verbal and quantitative measures using full-information factor analysis* (Report No. TM 860 575). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 275 698).

Kingston, N. M., & McKinley, R. L. (1988, April). *Assessing the structure of the GRE general test using confirmatory multidimensional Item Theory.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

Koch, W. R. (1983). *The analysis of dichotomous test data using nonmetric multidimensional scaling* (Report No. TM 830 617). Austin, TX: The University of Texas at Austin. (ERIC Document Reproduction Service No. ED 235 204).

Liou, M. (1988). Unidimensionality versus statistical accuracy: A note on Bejar's method for detecting dimensionality of achievement tests. *Applied Psychological Measurement, 12,* 381-386.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometrika Monograph No. 15, 32* (4, Pt. 2).

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics.* New York: McGraw-Hill.

Morgan, R. (1989, March). *An examination of the dimensional structure of the ATP biology achievement test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417-430.

Nandakumar, R. (1987). *Refinement of Stout's procedure for assessing latent trait dimensionality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

BEST COPY AVAILABLE

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses-Comparison of different approaches. *Journal of Educational Measurement, 31,* 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18,* 41-68.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reckase, M. D. (1981). *Guessing and dimensionality: The search for a unidimensional latent space* (Report No. TM 810 389). Columbia, MO: University of Missouri. (ERIC Document Reproduction Service No. ED 204 394).

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Rosenbaum, P. (1984). *Testing the local independence assumption in item response theory* (Technical Report No. 84-85). Princeton, NJ: Educational Testing Service.

Roussos, L., & Stout, W. F. (1994, April). *Analysis and assessment of test structure from the multidimensional perspective.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement, 15,* 109-127.

Steiger, J. H. (1980a). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Steiger, J. H. (1980b). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research, 15,* 335-352.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Stout, W. F., Junker, B., Nandakumar, R., Chang, H. H., & Steidinger, D. (1991). DIMTEST and TESTSIM [Computer programs]. Urbana, IL: University of Illinois, Department of Statistics.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, D., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific, Software, Inc.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293-308.

Zwick, R. W., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").